Human Pose Estimation with Rotated Geometric Blur

Bo Chen, Nhan Nguyen, and Greg Mori
School of Computing Science
Simon Fraser University
Burnaby, BC, CANADA
{bca11,tnn}@sfu.ca, mori@cs.sfu.ca

Abstract

We consider the problem of estimating the pose of a human figure in a single image. Our method uses an exemplar-matching framework, where a test image is matched to a database of exemplars upon which body joint positions have been marked. We find the best matching exemplar for a test image by employing a variant of an existing deformable template matching framework. A hierarchical correspondence process is developed to improve the efficiency of the existing framework. Quantitative results on the CMU MoBo dataset verify the effectiveness of our approach.

1. Introduction

In this paper we present a method for estimating the pose of a human figure. Our method attempts to automatically locate the 2D image positions of a set of keypoints (shoulders, elbows, etc.) on the human body. Figure 1 shows an example of the output of our method.

A robust solution to the problem of human pose estimation would have a variety of applications. The results can be used in gait analysis for either medical or surveillance applications. Pose estimation is needed to initialize kinematic trackers (e.g. Ramanan et al. [7]). Image synthesis applications in computer graphics (e.g. Hornung et al. [5]) can also benefit from automated techniques for pose estimation.

We develop a method which uses a database of labeled exemplars. In order to estimate the pose of a previously unseen person in a test image, the most similar exemplar image is found. This similar exemplar is found by matching sample points on the exemplar to sample points in the test image. The labeled positions of the body joints from the exemplar are then transferred onto the test image. It is well-known that this exemplar matching style of approach to pose estimation has difficulty in scaling to people in varying poses. In particular, human figures are articulated objects and the changes in the positions of limbs due to rota-
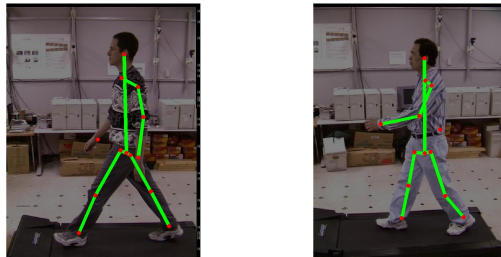

Figure 1. Sample results of pose estimation. Limb joints are marked with red dots and connected with green lines. The left arm is not visualized due to partial occlusion.

tion must be handled.

In this paper we apply the previously developed deformable template matching framework of Berg et al. [1], which uses *geometric blur* descriptors and integer quadratic programming (IQP) to solve for correspondences on pairs of shapes, to the problem of human pose estimation using exemplars. The main contribution of this paper is developing an application of this framework for human pose estimation. We develop a hierarchical approach to correspondence, and a refining sampling method for more efficiently obtaining dense correspondences between points on exemplars and points in test images. We show the effectiveness of this matching framework over previous approaches for exemplar-based human pose estimation.

The rest of this paper is organized as follows. First, we review previous work in Section 2. Our deformable template-based method for pose estimation is described in Section 3. We present experimental results in Section 4 and conclude in Section 5.

2. Previous Work

The problem of human pose estimation has been the subject of a vast amount of research in the computer vision community. Forsyth et al. [3] and Gavrila [4] provide comprehensive surveys of the related literature.

One line of work in this literature estimates pose by matching to stored exemplars using shape cues. Toyama and Blake [13] developed a probabilistic exemplar tracking model, and an algorithm for learning its parameters. Sullivan and Carlsson [11] and Mori and Malik [6] directly addressed the problem of pose estimation. They stored sets of 2D exemplars upon which joint locations have been marked. Joint locations were transferred to novel images using shape matching. Matching to similar exemplars was performed using *order structure* and *shape contexts* respectively. Shakhnarovich et al. [10] addressed variation in pose and appearance in exemplar matching through brute force by automatically constructing a large set of synthetic exemplars using computer graphics models. A variant of locality sensitive hashing was used for speed to match upper body configurations of standing, front facing people in background subtracted image sequences.

In our work we use geometric blur and integer quadratic programming (IQP) to operationalize the deformable template matching of exemplars to test images, based on the work of Berg et al. [1]. The cost function used in the IQP measures matching costs of geometric blur descriptors at sample points, in addition to pairwise distortion costs. In the IQP formulation, the structure of these pairwise costs can be arbitrary. Other related methods use tree-structured pairwise distortion costs. An example is the work of Thayananthan et al. [12], who added figural continuity constraints to shape context matching of contours for detecting hand poses. When the distortion costs are tree-structured, efficient distance transform matching algorithms [2] can be used, as was done for shape context matching by Rova et al. [9]. However, there are spatial constraints, for example between points on the right and left legs, or across a single limb, which are not captured in tree-structured pairwise distortion costs. Hence, we instead choose to use the IQP framework in our approach.

3. Approach

Our approach attempts to find the best matching exemplar image for a given test (query) image. We represent each exemplar image with a discrete set of points sampled from the image. For each exemplar image, we attempt to deform it into alignment with the test image. The exemplar image which best matches the test image will be used to estimate the pose of the figure, by transferring the labeled joint positions from exemplar to the positions given by alignment with the test image.

A cost function is defined to evaluate the matching of two point sets P and Q sampled from the exemplar image and the query image respectively. The cost function consists of two terms: the first term, which measures the similarity between putatively matched points $p_i \in P$ and $q_{i'} \in Q$, is based upon the geometric blur descriptor with rotation.

The second term measures distortion costs between pairs of points $p_i, p_j \in P$ and $q_{i'}, q_{j'} \in Q$.

3.1. Geometric Blur

In order to determine the matching between an exemplar image and a test image, we employ the geometric blur(GB) descriptor of Berg et al. [1]. The GB descriptor is a point descriptor which captures the coarse arrangement of gradient information relative to the location of the point in question.

Oriented gradient filter responses (channels) from the image are computed in a variety of directions (we use horizontal, 45°, vertical, and 135° in our experiments). Absolute values are taken for each channel to lessen the influence from clothing. For instance, the channel of a person in black shirt on white background should be considered similar in terms of shape to the channel of the same person wearing white shirt on black background. The channels are then compressed by a sigmoid function in order to reduce the effects of high contrast boundaries.

These channels are then blurred with a spatially varying blur kernel to increase robustness to small shifts in gradient positions. As in [1], we define $S_d = S * G_d$, where our original channel signal S is convolved with G_d, a Gaussian kernel of width d. The geometric blur descriptor for a sample point x_0 is then

$$B_{x_0}(x) = S_{\alpha|x|+\beta}(x_0 - x) \qquad (1)$$

where α and β are constants which control the spatial variation in blur. The geometric blur is computed for all points of interest. For each of the 4 channels, we sample along 12 equally spaced orientations at 4 logarithmically spaced radii. This $4 \times 12 \times 4$-dimension vector gives the GB descriptor. $L2$ distance between descriptors is used as the geometric blur matching cost.

In the presence of limb rotations, two points at similar locations may not be well correlated by the L2 distance. Nevertheless if the angle of rotation is known in advance, one descriptor can be pre-rotated accordingly to align with the other, as illustrated in fig 2. Automatic estimation of the rotation angle is discussed in section 3.3.2.

For consistency in notation we use $GBC_\theta(p_i, q_{i'})$ to denote the GB matching cost between the GB descriptor of p_i under θ-degree rotation, and the descriptor of $q_{i'}$. The GB cost in [1] would be a special case where $\theta = 0$.

3.2. Distortion Costs

To quantify the pairwise consistency of a correspondence matching $p_i, p_j \in P$ to $q_{i'}, q_{j'} \in Q$. We first compute the offset vector $\vec{r} = p_j - p_i$ and $\vec{s} = q_{j'} - q_{i'}$, then evaluate the following distortion function $D(\vec{r}, \vec{s}) = \gamma C_{rot}(\vec{r}, \vec{s}) +$

Figure 2. Visualization of geometric blur descriptors. For each descriptor, three inner radii of the second channel are shown as gray-scaled dots, where darker dots correspond to greater channel responses. (a) and (b) are both computed at a point slightly below the knee, but the rotation angles of two legs differ by $\frac{\pi}{3}$. (c) is a rotated version of (a) by $\frac{\pi}{3}$. (b) is closer to (c) than (a) in terms of L2 distance.

$(1 - \gamma)C_{sc}(\vec{r}, \vec{s})$, where:

$$C_{rot}(\vec{r}, \vec{s}) = (\frac{\alpha_d}{|\vec{r}||\vec{s}|} + \beta_d)|arccos(\frac{\vec{r} \cdot \vec{s}}{|\vec{r}||\vec{s}|})| \quad (2)$$

$$C_{sc}(\vec{r}, \vec{s}) = \frac{||\vec{r}| - |\vec{s}||}{|\vec{r}| + |\vec{s}|} \quad (3)$$

C_{rot} corresponds to the rotational cost and C_{sc} is the scaling cost. The weighted sum of these costs by a parameter γ gives the total distortion cost. The scale and variance of the rotational cost is controlled by constant parameters α_d and β_d. In the experiments we used $\gamma = 0.1$ to give some amount of tolerance in rotation. We also used $\alpha_d = 10$ and $\beta_d = 1$.

There are some changes in our distortion function as compared to the one in [1]. The distortion function used in [1] is:

$$\hat{C}_{rot}(\vec{r}, \vec{s}) = (\frac{\alpha_d}{|\vec{r}|} + \beta_d)|arcsin(\frac{\vec{r} \times \vec{s}}{|\vec{r}||\vec{s}|})| \quad (4)$$

$$\hat{C}_{sc}(\vec{r}, \vec{s}) = \frac{|\vec{r}| - |\vec{s}|}{|\vec{r}| + \mu_d} \quad (5)$$

where μ_d is a balancing parameter.

First, we use arccos to assign a high cost to the degenerate case where $|\vec{s}| = 0$, i.e. multiple points from the exemplar collapsed into one matched point in the query. Second, in the scaling cost we take the absolute value to avoid negative costs. Finally, we enforce the cost to be symmetric ($D(\vec{r}, \vec{s}) = D(\vec{s}, \vec{r})$).

3.3. Correspondence

For the exemplar image and query image we construct P and Q respectively by sampling evenly on the edge map obtained from Canny edge detection and background subtraction. To avoid omission of points, the query image is more densely sampled than the exemplar image. For efficiency, instead of computing all possible correspondences, each point $p_i \in P$ is restricted to match to a smaller set $\hat{Q}(i)$ of m putative good matches.

A correspondence $\delta \in \{0, 1\}^n$ maps each of the $|P|$ exemplar points to one of m best matches($n = |P| \times m$). $\forall p_i \in P, q_j \in \hat{Q}(i)$, the binary indicator $\delta_{ij} = 1$ iff p_i is matched to q_j.

The matching quality is measured by the weighted sum of the geometric blur matching cost and the distortion cost:

$$cost(\delta) = \alpha \cdot C_{match}(\delta) + (1 - \alpha) \cdot C_{dist}(\delta) \quad (6)$$

$$C_{match}(\delta) = \Sigma_{i,j:\delta_{ij}=1} GBC_0(p_i, q_j) \quad (7)$$

$$C_{dist}(\delta) = \Sigma_{i,i',j,j':\delta_{ii'}=\delta_{jj'}=1} D(j - i, j' - i') \quad (8)$$

$0 \le \alpha \le 1$ weighs the pointwise geometric blur cost and the pairwise distortion cost.

Berg et al. use an $O(n^2)$ approximation algorithm for solving this IQP. However, for our problem the number of sample points is much larger than that in [1]. The typical value for product of the number of sample points in the exemplar and putative good matches in the query is $n = 250 \times 50$. In addition, local adjustments in this IQP affects the solution globally, which is inefficient for methods making frequent local changes.

We develop a method of finding correspondence in two main steps. The first step identifies potential regions for good matching, and the second step searches for a finer global correspondence as well as the local limb rotations.

3.3.1 Seeding

In this section we describe our method for obtaining potential good matching sets $\hat{Q}(i)$ for each exemplar sample point. We refer to the procedure as "seeding", during which a rough correspondence is estimated to obtain a set of potential regions. Each region R_q is a circle with radius r centered at a representative query point q.

Seeding does not require an accurate, detailed correspondence, therefore it is sufficient to use only an initializer set $P_0 \subset P$, which consists of sparsely distributed points.

For each $p_i \in P_0$ we keep the top m points in Q with smallest geometric blur cost with respect to p_i. Then the best correspondence of P_0 is obtained by solving an IQP. The result δ_0 maps P_0 to Q_0, the set of representative query points.

Finally, for each $p_i \in P$, the set of m best matches $\hat{Q}(i)$ is chosen from $R = \cup_{\forall q_{i'} \in Q_0} R_{q_{i'}}$.

In the experiment we use $|P_0| = 100$ and $m = 15$, reducing the running time of IQP to roughly 1/50.

A sample result of seeding is shown in Fig. 3.

3.3.2 Refined Correspondence

Seeding does not provide us highly reliable correspondence. Refinement based on the current coarse matching and exploiting the human body configuration is now employed.
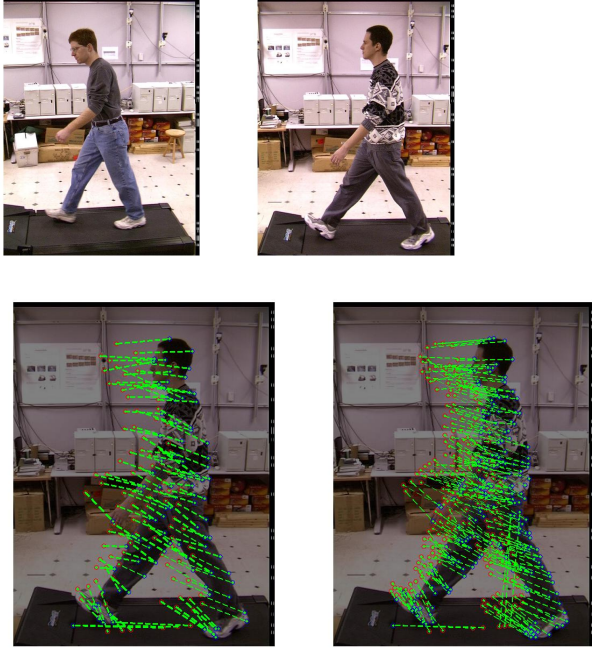
Figure 3. The top row shows the exemplar image and the query image. Bottom left: seeding; Bottom right: the correspondence. Exemplar points are marked in red, query points in blue, and the matchings are shown in green lines.

In specific, consider the tree-model with four chains of deformation $\mathbb{K}_i$, $i = 1..4$, where $\mathbb{K}_1 = \{\text{torso, left arm}\}$; $\mathbb{K}_2 = \{\text{torso,right arm}\}$; $\mathbb{K}_3 = \{\text{torso, left thigh, left leg}\}$ and $\mathbb{K}_4 = \{\text{torso, right thigh, right leg}\}$.

We assume that each body part is rigid enough such that rotations for all its points are uniform.

Our objective is solving simultaneously for $\delta = \{\delta_b\}$, the optimal matching and $\Theta = \{\theta_b\} \in [-\pi, \pi]^9$, the optimal rotation, for each of the 9 body parts along the chain. Here we assume that each limb is rigid enough such that rotation angles of its points are uniform.

The matching cost is:

$$cost(\delta, \Theta) = \alpha C_{match}(\delta) + (1 - \alpha)C_{dist}(\delta) \quad (9)$$

$$C_{match}(\delta, \Theta) = \sum_{b=1}^{9} GBC_{\theta_b}(\delta_b); \quad (10)$$

$$C_{dist}(\delta) = \sum_{b=1}^{9} D(\delta_b, \delta_b)$$

$$+ \sum_{i=1}^{4} \sum_{b_1 \neq b_2 \in \mathbb{K}_i} D(\delta_{b_1}, \delta_{b_2}) \quad (11)$$

where $D(\delta_x, \delta_y)$ corresponds to the distortion cost of the matching of body part y given the matching of part x. When $x \neq y$, $D(\delta_x, \delta_y)$ represents the total part-wise distortion

cost between part x and part y, which depicts the consistency constraint enforced within a chain.

We approximate the problem by discretizing the angles of rotation, and assuming only forward constrain propagation along the deformation chain:

Initial Step: Since torso is the common root for all chains and it typically is subjected to little rotation, an intuitive idea is to set $\theta_{torso} = 0$ and independently solve for δ_{torso}. This is again an IQP optimization with smaller scale, for which only torso points with their top m best matches in terms of geometric blur cost are considered.

Iterative Step:

- After b_{i-1}, the $(i-1)^{th}$ part along chain $\mathbb{K}$, has been matched $(1 < i \leq length(\mathbb{K}))$, consider b_i, the next part along $\mathbb{K}$. Since $\forall j < i, \delta_{b_j}$ is known, the consistency imposed by the prior matching on part b_i, denoted by $Cs(\delta_{b_i})$, is pre-computable, and hence a linear constraint.

- We search for the best rotation angle θ_{b_i} by iteratively attempting a set of discretized values. For each attempted θ_{b_i}, the m best matches for each point in part b are recomputed from the seeded region R as in section[3.3.1].

- The objective function for part b_i has been simplified to

$$
\begin{aligned}
cost(\delta_{b_i}) = & \; min_\theta \{\alpha \times GBC_\theta(\delta_{b_i}) \\
& + (1 - \alpha)Cs(\delta_{b_i}) + (1 - \alpha)D(\delta_{b_i}, \delta_{b_i})\}
\end{aligned}
$$

which defines an optimization problem that can be solved locally along with the corresponding best θ.

This procedure can be carried out recursively to solve for all parts along the chain. Sample results are shown in Fig.3 Although it might appear to be a lengthly process, in practice the range of rotation can be further restricted, and each iteration involves solving IQP for only a small set of points belonging to the part($|X_p| \simeq 20$), thus significantly reducing the number of calculations.

3.4. Deformation

Given the correspondences obtained via the method above, we estimate a deformation to transform the best exemplar into alignment with the test image.

The least square best transform for each body part is estimated, as in [6]. To eliminate outliers without presuming a fixed probability of inliers, RANSAC is applied in an iterative manner, i.e. iteratively relax the RANSAC requirement for assertion after a certain number of estimated transforms all fail to fit the model. The procedure terminates when either a suitable transformation is found or the requirement drops below certain threshold value.

Finally, we deform each joint point of the exemplar according to the estimated transformation. The deformed joint locations are the resultant estimation of the pose of the figure in the test image.

4. Experiments

To evaluate our method's ability to handle variance in shape, clothing, and visibility, experiments are performed on the side-view fast walk `fastWalk/vr03_7` dataset from CMU Mobo database [8]. 8 sets of subjects with 30 frames each are selected as our test sets. 10 evenly distributed frames from each test set (8×10 frames) are selected as the query set. For each query image, 8 randomly selected frames from other test sets (7×8 frames) are used as exemplars, which are matched to the query image using our method. The joint locations of the exemplar are manually marked, which are then deformed to estimate the joint positions of the query image according to the sample point correspondence. Our unoptimized implementation in MATLAB estimates the matching from each exemplar to the query image in 5 minutes.

The sample results are visualized in Fig. 4 . We color the visible joints and connect the points to form the torso, the right arm and both legs.

As shown in the figure, joint locations for non-occluded joints are successfully captured by deformable matching. Also our method has more robustness against clothing texture primarily because the independent chains of deformation prevent the local, bad correspondence from spreading.

Difficulty does arise, however, when estimating limbs that are occluded, under severe rotation,or has insufficient points or edges due to inaccurate background subtraction.

In the occlusion case, the geometric blur signals from multiple limbs interfere each other so that rotation alone is not enough to correlate non-overlapping joints with the overlapped ones, and our assumption of independent deformation does not hold. In this case, an exemplar frame that has the similar occlusion of joints are required to build good correspondence.

The full results are compared to Mori and Malik [6] in Table 1. Our approach outperforms the shape context matching method in two aspects: First, more precise joint positions are estimated using less frames (only 8 frames versus 30 frames used in the shape context approach) for each exemplar set. Second, our method are less sensitive to clothing and rotational variance.

5. Conclusion

In this paper we have presented an application of an existing deformable template matching framework to the problem of human pose estimation. The geometric blur and integer quadratic programming framework was modified to include a more efficient hierarchical search strategy for computing correspondence. This framework was demonstrated to be effective, obtaining joint position error rates on the CMU MoBo dataset which were significantly lower than previous approaches while using fewer exemplars. This is due to our method's improved ability to handle rotation of parts, and the cost function which is able to model more complex dependencies between sample points.

References

[1] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 26–33, 2005.

[2] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical Report TR2004-1963, Cornell Computing and Information Science, 2004.

[3] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion: Part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1(2), 2006.

[4] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.

[5] A. Hornung, E. Dekkers, and L. Kobbelt. Character animation from 2d pictures and 3d motion data. *ACM Trans. Graph.*, 26(1), 2007.

[6] G. Mori and J. Malik. Recovering 3d human body configurations using shape context matching. *IEEE Trans. PAMI*, 28(7):1052–1062, 2006.

[7] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2005.

[8] R.Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001.

[9] A. Rova, G. Mori, and L. M. Dill. One fish, two fish, butterfish, trumpeter: Recognizing fish in underwater video. In *IAPR Conference on Machine Vision Applications*, 2007.

[10] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *Proc. 9th Int. Conf. Computer Vision*, volume 2, pages 750–757, 2003.

[11] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *European Conference on Computer Vision LNCS 2352*, volume 1, pages 629–644, 2002.

[12] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume I, pages 127–133, Madison, USA, June 2003.

[13] K. Toyama and A. Blake. Probabilistic exemplar-based tracking in a metric space. In *Proc. 8th Int. Conf. Computer Vision*, volume 2, pages 50–57, 2001.
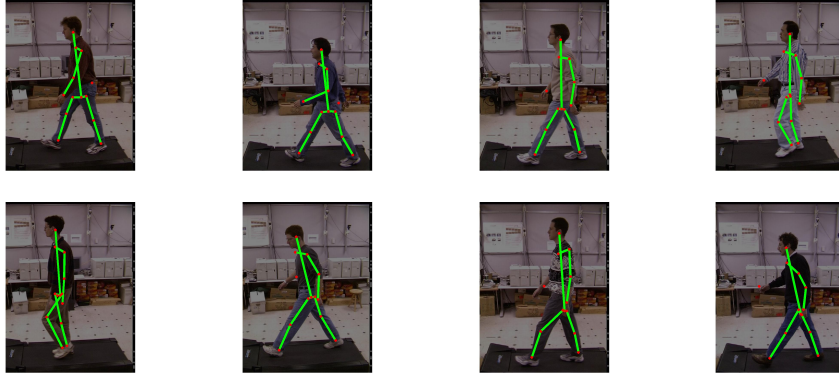
Figure 4. Sample results. Joint points of the left arm and the legs are marked in red and connected with green lines. The right arm is not always entirely visible, hence it is not visualized.

	Shoulder	Elbow	Hand	Hip	Knee	Ankle
	7.9 ± 5.0	15.9 ± 6.0	12.4 ± 8.1	11.5 ± 4.6	10.7 ± 5.1	4.4 ± 1.9
	12.9 ± 9	27.0 ± 26	43.3 ± 47	18.1 ± 11	32.9 ± 36	45.8 ± 59
	10.1 ± 3.5	28.2 ± 13.0	23.6 ± 18.0	18.0 ± 9.0	12.6 ± 8.1	3.2 ± 1.8
	15.1 ± 7	18.4 ± 9	26.5 ± 15	16.6 ± 7	16.4 ± 10	16.4 ± 12
	7.6 ± 2.6	15.1 ± 5.7	13.1 ± 10.3	9.9 ± 2.8	12.3 ± 10.7	6.7 ± 3.4
	12.9 ± 13	22.3 ± 20	27.8 ± 20	15.2 ± 7	13.6 ± 7	17.4 ± 22
	10.4 ± 3.7	12.7 ± 5.0	9.8 ± 6.5	14.4 ± 5.9	14.6 ± 10.2	10.2 ± 13.4
	13.8 ± 8	29.4 ± 16	27.4 ± 25	22.8 ± 14	24.4 ± 25	24.1 ± 23
	8.4 ± 1.6	18.3 ± 9.6	11.1 ± 8.5	10.9 ± 6.5	18.1 ± 18.4	7.4 ± 3.4
	42.8 ± 24	69.1 ± 40	98.7 ± 54	41.8 ± 23	56.1 ± 27	79.0 ± 55
	8.3 ± 3.1	15.3 ± 7.0	13.7 ± 12.0	13.8 ± 10.9	14.3 ± 7.8	12.9 ± 11.9
	13.8 ± 9	31.9 ± 17	34.7 ± 30	16.1 ± 8	19.6 ± 20	26.8 ± 42
	11.7 ± 3.2	15.3 ± 5.8	27.3 ± 17.3	10.8 ± 7.4	14.8 ± 11.7	7.3 ± 4.2
	13.2 ± 8	24.0 ± 10	27.7 ± 37	15.6 ± 9	21.0 ± 28	24.7 ± 48
	15.2 ± 6.1	16.4 ± 5.7	16.1 ± 10.1	25.4 ± 9.3	15.7 ± 13.3	10.7 ± 4.3
	17.9 ± 11	34.4 ± 33	45.8 ± 41	25.7 ± 12	34.6 ± 29	45.9 ± 50
	24.9 ± 16.3	24.3 ± 21.1	11.3 ± 14.0	7.6 ± 6.6	9.0 ± 5.1	10.0 ± 5.4
	17.9 ± 8	22.6 ± 13	28.5 ± 24	16.0 ± 9	20.7 ± 20	23.5 ± 37
Mean	11.6	17.9	15.4	13.6	13.5	8.1
Mean	17.8	31.0	40.0	20.9	26.5	33.7

Table 1. Mobo database "fast walk side-view" subject numbers versus joint position error. The error is based on the distance of computed position from ground truth in pixels. Each cell shows error mean and standard deviation by our proposed method (top) and that of shape context exemplars [6] (bottom). The last row shows the mean error over all 9 subjects in pixels.